

# Role of Scene text in Image Semantics

*A Thesis submitted by*  
**Arka Ujjal Dey**

*in partial fulfillment of the requirements for the award of the degree of*  
**Doctor of Philosophy**



॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

**Indian Institute of Technology Jodhpur**  
**Department of Computer Science and Engineering**

*August 2022*



## Declaration

I hereby declare that the work presented in this Thesis titled *Role of Scenetext in Image semantics* submitted to the Indian Institute of Technology Jodhpur in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out under the supervision of Dr. Gaurav Harit. The contents of this thesis in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

*Arka Ujjal Dey*  
*P15CS001*



## Certificate

This is to certify that the thesis titled *Role of Scenetext in Image Semantics*, submitted by *Arka Ujjal Dey (P15CS001)* to the Indian Institute of Technology Jodhpur for the award of the degree of *Doctor of Philosophy*, is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

*Gaurav Harit*  
*Ph.D. Thesis Supervisor*



## Acknowledgments

I thank Asimov and ELIZA for making me think how similar we are to automatons. And to my family and their appetite for academics, making me believe research is 'cool'. This work would not have been possible without Dr.Harit, those sustained support has helped it take shape. A special mention to my colleagues, in IIT Jodhpur and CVC Barcelona, for the healthy debates.





## List of Figures

<i>Figure</i>	<i>Title</i>	<i>Page</i>
1.1	Contextual Encoding– Apply attentional framework on a graph-structured organization of detected visual and scene text objects to generate contextual encoding. The said encoding is applied to downstream tasks like classification, retrieval, and even question answering.	2
1.2	Knowledge scheme – Use context validated external knowledge facts along with detected visual and scene text objects to answer questions about an image.	4
2.1	Complementary nature of the text and visual cues: In some cases, the visuals can be symbolic, but the embedded text gives away the context[top-left, top-right], in other cases, the visuals can be simple to understand, but the text can be obtuse[bottom-left]. Further, the amount of text content can vary widely [top-right, bottom-right]	10
3.1	Contextual Embedding scheme – Use detected visual symbolism and objects, together with scene text to reason about images.	21
3.2	Model architecture of the Proposed Contextual Embedding applied to the separate tasks of semantic embedding and classification.	22
3.3	Anchor based text encoding allows us to retain scene text information only relevant to the final task. Given an image and 5 queries statements to rank (first 2 correct, next 3 incorrect), we propose using the queries itself to filter scene texts. The cells correspond to similarity between current word and anchor given by $r_{n,k}$ in eq. 3.1 Dark values correspond to scene text that are relevant to the query or anchor and the rest of the scene text, marked white, is ignored.	24
3.4	Attention Mechanism: We use the context or neighbourhood of a node to define its rich contextual representation. For example for node $x_1$ , all its neighbouring nodes $x_2, x_3, x_4, x_5, x_6$ along with its initial representation $x_1$ contributes to the final representation $h_1$ , weighted by their relevance indicated by $\alpha_{i,j}$	25
3.5	A sample advertisement image, with relevant sentences in blue and irrelevant sentences in red. The task is to rank the relevant sentences ahead of the irrelevant ones. Showing only 5 of 15 statements for brevity	26
3.6	Retrieval Framework: The Action and Reason semantic embedding encodes the image and scene text content, but since the final task is about ranking query statements we also compare the scene text with the query statements in terms of semantics and lexical similarity	27
3.7	Sentence Relevance Task: Training, Testing. The Separate text channels allows us to account for scene text words that may have missed by the contextual encoder.	28
3.8	Semantic Retrieval. The query image is at the top left, with a red border. The rest of the top row images show the best matches retrieved using semantic features based on visual cues. The bottom row shows the results obtained by incorporating the contextual encoder for semantic features using both scene text and visual cues. Our improvements lead to the retrieval of images, not just about cars but also getting the type and brand right.	32
3.9	Ablated Instances	32
3.10	Qualitative Results on the Classification task. Correct class labels are in green, and incorrect ones are in red	34

4.1	Knowledge scheme – Use context validated external knowledge fact, marked blue, along with detected visual and scene text objects to answer questions about an image.	37
4.2	Extraction, Validation, and Reasoning: The three stages of using external knowledge in Text-VQA. Extraction corresponds to query GKb with scene text tokens and retrieving candidate knowledge facts, which in stage 2 is validated with context, and finally in stage 3, this context validate knowledge is incorporated in to reasoning during answer generation	41
4.3	Knowledge Extraction and Validation. Figure (a) shows Candidate knowledge fact extraction using Google API, and Figure (b) demonstrates knowledge validation and selection through image context. We identify knowledge validation as an important subtask for web scraped knowledge data that is not anchored to our Text-VQA datasets explicitly	42
4.4	Constrained Interaction ensures that a scene text token only interacts with its own knowledge fact. The transformer Attentional framework allows complete interaction between all question words (blue highlighted) visual (green boxes), scene text (yellow boxes) and retrieved knowledge facts( orange with highlighted black). However as the answer words are selected from the scene text tokens we ensure that only the legitimate scene text-knowledge fact pairs interact through constrained interaction. For example, in the sample above, the knowledge fact about Carnegie hall should only interact with 'carnegie hall' and no other scene text.	45
4.5	Constrained Interaction: Knowledge inclusion through Attention Mask matrix $A$ . It controls the interaction between the nodes. Dark Cells allow interaction between corresponding nodes. $A^{KT}$ and $A^{TK}$ , highlighted in blue, are $N \times N$ identity matrices, binding the knowledge facts with their corresponding OCR tokens. Highlighted in pink is the $D \times D$ lower triangular sub-matrix $A^{PP}$ enforcing causality between previous decoding steps	46
4.6	EKTVQA: Our proposed External Knowledge-enabled Text-VQA	47
4.7	Examples of knowledge effectiveness. TVQA is our baseline similar to M4C [Hu <i>et al.</i> , 2020] and does not use external knowledge. The correct answers are marked in Blue, while incorrect ones are marked in Red. Multiword Entities, row 3, are marked with Purple. In Row 1, we observe that we can improve answer entity correctness with knowledge. Row 2 shows the effectiveness of external knowledge in dealing with Dataset Bias. Finally, in Row 3, we showcase our Multi-word Entities.	52
4.8	Examples of knowledge limitations. The correct answers are marked in Blue, while incorrect ones are marked in Red. Multiword Entities are marked with Purple	53
5.1	Scene text Reference. Highlighted in red are the scene text contents referenced in the questions. Marked yellow are the answer scene text tokens. To be noted is the fact that the answer scene text tokens are near the question-referred scene text tokens. The two images on the right are annotated with the scene text token for visual clarity.	60
5.2	Training scheme for consistency: Similar questions must look at similar visual and scene text cues.	61
5.3	Rephrased question examples. Highlighted in yellow are the rephrased versions of the questions above.	62
5.4	Attention scores that determine the answer decoding. To the left, we unroll the transformer layers, depicting, in particular, the attention scores of the final decoding step. $\alpha_z^{prv}$ is the weighing factor for the rest of the nodes in the final representation of the decoder input $\mathbf{z}^{prv}$	63
5.5	Dependence of answers on scene text. In highlighted boxes, we demonstrate the possible types of questions ignored because they do not depend on scene text	64

## List of Tables

<i>Table</i>	<i>Title</i>	<i>Page</i>
2.1	Details of Publicly available Dataset with Scene text content	13
2.2	Comparison of Text Representation schemes	18
3.1	Comparison with state-of-the-art. Results marked with * do not use exactly our same partitions for training and test.	30
3.2	Semantic Embedding: Role of Text and Visual channels, partitioning and Contextual VT encoder in semantic embedding	31
3.3	Sentence Relevance: Role of components Text Semantic, Text Lexical and Contextual Embedding in Sentence Relevance task	31
3.4	Topic Classification results	33
3.5	Computation cost : O: Visual Object, S: Visual Symbolism, T: Scene Text, BU: Bottom-Up, CoAt: Co-Attention, GAT: Graph Attention, LS: LSTM for statement encoding, HN: Hard Negative mining for Triplet, KB: Knowledge Branch training, TS: Cosine Distance based Text scoring, n: number of objects in an image (text visual or scene text), $d_1$ : feature input dimension, $d_2$ : feature output dimension *Approximate detection cost based on components used	35
4.1	External knowledge related work summary	39
4.2	Results of Text-VQA Task on TextVQA and ST-VQA Dataset	48
4.3	Results of Text-VQA Task on Text-KVQA (Scene) Dataset. Results marked with * do not use exactly our same partitions for training and test.	48
4.4	Dataset Comparison. <b>OCR UB</b> : Percentage of answers comprising OCR tokens alone, <b>Vocab UB</b> : Percentage of answers comprising fixed vocabulary words only, <b>OCR + Vocab UB</b> : Percentage of answers comprising OCR tokens and answer vocabulary words, <i>i.e.</i> , the maximum achievable accuracy.	50
4.5	Results of related methods on TextVQA Dataset with additional pre-training tasks and datasets	51
4.6	Upper Bounds and OCR systems: <b>OCR UB</b> : Percentage of answers comprising OCR tokens alone, <b>Vocab UB</b> : Percentage of answers comprising fixed vocabulary words only, <b>OCR + Vocab UB</b> : Percentage of answers comprising OCR tokens and answer vocabulary words, <i>i.e.</i> , the maximum achievable accuracy.	53
4.7	Ablation study 1: Role of Knowledge related Components	54
4.8	Ablation study 2: Transferable property of external knowledge	56
5.2	Upper Bounds and OCR systems: <b>OCR UB</b> : Percentage of questions answerable with OCR tokens alone, <b>Vocab UB</b> : Percentage of questions answerable using fixed vocabulary words, <b>OCR + Vocab UB</b> : Percentage of questions answerable using OCR tokens and fixed vocabulary words, <i>i.e.</i> , the maximum achievable accuracy.	65
5.3	Role of Scene Text and Visual Components In Text-VQA. Scene text cues dominate.	65
5.4	Sentence Relevance: Role of Scene text and Visual Components in Sentence Relevance task. Text cues are lead to better results than visual cues.	65



## List of Symbols

$A$	Attention Matrix
$\mathbf{r}_{i,k}$	gives the similarity between a scene text word $t'_i$ and anchor $A_k$
$\mathbf{t}'_i$	scene text feature for $i^{th}$ word
$\mathbf{t}_k$	scene text features in terms of $k^{th}$ anchor
$\mathbf{v}_i^{res}$	ResNet-152 Convolutional Features for the $i^{th}$ object
$\mathbf{v}_i$	Visual Features for the $i^{th}$ object
$W$	learned projection matrix
$\alpha_{ij}$	the attention weight representing the importance of node $j$ to node $i$
$\alpha_z^{prv}$	The weighing factor for the rest of the nodes in the final representation of the decoder input
$d_{cs}$	Refers to the cosine distance
$d_L$	Refers to the edit distance
$D$	Number of Decoding Steps
$H$	Number of Vocabulary objects in classifier
$L$	Number of Question Words
$M$	Number of Visual Objects
$N$	Number of Scene text Objects
$\mathbf{r}_j$	validity score for $j^{th}$ candidate knowledge $\mathbf{S}_j \in \mathbf{S}_i$
$\mathbf{S}_i$	Set of candidate knowledge facts for scene text word $\mathbf{t}_i$
$\mathbf{x}_p^c$	Feature representation of $p^{th}$ context object
$\mathbf{x}_m^{obj}$	Feature representation of $m^{th}$ visual object
$\mathbf{x}_n^{ocr}$	Feature representation of $n^{th}$ scene text object
$\mathbf{x}_n^{knw}$	Feature representation of the knowledge fact corresponding to $n^{th}$ scene text object
$\mathbf{x}_d^{prv}$	Feature representation at the $d^{th}$ step of the previous-step object
$\mathbf{x}_l^{ques}$	Feature representation of $l^{th}$ question word
$\mathbf{y}_d^{ocr}$	Pointer network score for OCR objects at the $d^{th}$ step
$\mathbf{y}_d^{voc}$	classifier score for answer vocabulary objects at the $d^{th}$ step
$\mathbf{z}_m^{obj}$	Transformer output of $m^{th}$ visual object
$\mathbf{z}_n^{ocr}$	Transformer output of $n^{th}$ scene text object
$\mathbf{z}_n^{knw}$	Transformer output of the knowledge fact corresponding to $n^{th}$ scene text object
$\mathbf{z}_d^{prv}$	Transformer output of $d^{th}$ previous decoding step
$\mathbf{z}_l^{ques}$	Transformer output of $l^{th}$ question word



## List of Abbreviations

<b>ANLS</b>	Averaged Normalized Levenshtein Similarity
<b>BERT</b>	Bidirectional Encoder Representations from Transformers
<b>BiLM</b>	Bi-directional Language Model
<b>BOW</b>	Bag Of Words
<b>CE</b>	Cross Entropy
<b>CNN</b>	Convolutional Neural Network
<b>CVC</b>	Computer Vision Center
<b>EKB</b>	External Knowledge Base
<b>EKTVQA</b>	External Knowledge-enabled Text Visual Question Answering
<b>FC</b>	Fully Connected
<b>FCNN</b>	Fully Connected Neural Network
<b>F-RCNN</b>	Faster R-CNN
<b>GAN</b>	Generative Adversarial Network
<b>GAT</b>	Graph Attention Layers
<b>GKB</b>	Google Knowledge Base
<b>GCN</b>	Graph Convolution Network
<b>GPU</b>	Graphics Processing Unit
<b>ILSVRC</b>	ImageNet Large Scale Visual Recognition Challenge
<b>ITM</b>	Image Text (contrastive) Matching
<b>LSTM</b>	Long Short Term Memory
<b>M4C</b>	Multi-modal Multi-Copy Mesh
<b>MLM</b>	Masked Language Modeling
<b>MMHS150K</b>	Multimedia Hate Speech Dataset
<b>MSE</b>	Mean Squared Error
<b>NLP</b>	Natural Language Processing
<b>NLU</b>	Natural Language Understanding
<b>NLS</b>	Normalized Levenshtein Similarity
<b>OCR</b>	Optical Character Recognition
<b>PHOC</b>	Pyramid Histogram of Characters
<b>R@K</b>	Recall at K
<b>RPP</b>	Relative (spatial) Position Prediction
<b>ReLU</b>	Rectified Linear Unit
<b>SA-M4C</b>	Spatially Aware M4C
<b>SGD</b>	Stochastic Gradient Descent
<b>SMA</b>	Structured Multi-modal Attention
<b>SVM</b>	Support Vector Machine
<b>TAP</b>	Text aware Pretraining
<b>Text-VQA</b>	Text Visual Question Answering
<b>Text-KVQA</b>	Text Knowledge Visual Question Answering
<b>TF-IDF</b>	Term Frequency-Inverse Document Frequency
<b>TVQA</b>	Our baseline model for Text-VQA
<b>Vocab UB</b>	Vocabulary Upper bound referred
<b>VLN</b>	Vision and Language Navigation

**VQA** Visual Question Answering  
**VGG** Visual Geometry Group